Human Factors in Cybersecurity, Vol. 53, 2022, 36–43
https://doi.org/10.54941/ahfe1002199

Social Engineering and Human-Robot Interactions' Risks

Ilenia Mercuri

DeepCyber, Rome, RM, Italy

ABSTRACT

Social Robots are created in order to interact with human beings. This paper aims to provide insights into how the interaction with social robots could be exploited by humans not only in a positive way but also by using the same techniques of social engineering borrowed from "*bad actors*" or hackers, to achieve malevolent and harmful purposes for mankind. The Human Factor is the weakest ring of the security chain. There is a fine line that separates the opinions of those who argue that, in the future, machines with artificial intelligence could be a valuable aid to humans to those who believe that they represent a huge risk that could endanger human protection systems and safety. It is necessary to examine in depth this new field of cybersecurity to analyze the best path to protect our future. Are social robots a real danger?

Keywords: Human factors, Cybersecurity, Cyberpsychology, Social engineering attacks, Human-robot interaction, Robotics, Malicious artificial intelligence, Affective computing, Cyber threats

INTRODUCTION

The word "*robot*" comes from the Czech language and means "*forced labor*". It was first used by author Karel Capek in a 1920 play, "*Rossum's Universal Robots*". Over time, its meaning has become broader, indicating any type of machine capable of carrying out a job independently of humans. The oldest robot in history built with a human appearance was designed in the 3rd century B.C. by an engineer in Byzantium and is known today as the "*automatic servant of Philon*". It could serve wine and water to guests attending banquets. Throughout the centuries, humans have always sought to build faithful "*copies*" of themselves, able to perform a series of automated tasks in their stead. Think of automata, home appliances, computers, assistants of various kinds and - more recently - the growing development in the use of "*Artificial Intelligence*" for the construction of technologies of various kinds, which are becoming increasingly autonomous. Modern robotics seems to have taken root from the theories of another author, however: Isaac Asimov, in 1941, from his collection of short stories "*I, Robot*". Asimov proves to have been the author of what famously came to be known as the "*Three Laws of Robotics*". Considering these historical references, it is clear how robots have always been conceived in close connection with mankind, both in a positive sense (from a physical and social point of view), as well as in a negative sense (otherwise, it would not have been necessary to devise laws to protect the "*father*" of robots, man himself). Uniquely, "*social robots*" are created to interact with human beings; they have been designed and programmed to engage with people by leveraging a "*human*" aspect and various interaction channels, such as speech or non-verbal communication. They therefore readily solicit social responsiveness in people who often attribute human qualities to the robot. Social robots exploit the human propensity for anthropomorphism. In order to make the human/robot relationship as real as possible and similar to normal human interactions, we have long been curious about the characteristics that make humans human and robots not human.

ARTIFICIAL INTELLIGENCE, SOCIAL ROBOTS, AND SOCIAL ENGINEERING

One area of research that has become increasingly popular in recent decades is the study of "*Artificial Intelligence*" or "*A.I.*", which aims to use machines to solve problems that, according to current opinion, require intelligence. Luminaries such as Elon Musk and Dr. Stephen Hawking have recently stated that artificial intelligence could be the technology that will bring about the end of the human race. Recently, an artificial intelligence app called Replika was created. It has already been downloaded by more than seven million users worldwide and has already revealed some disturbing implications. Specifically, it is a "*chatbot*", i.e., a "*bot*" created to chat and equipped with the typical skills of what is described as "*affective computing*" (a concept better explained later in this work). The app was created to keep people company by posing as a friend, a lover, or a mentor, of either male or female gender. According to its developers, it could help improve personal mental health by providing "*psychological assistance*", although the terms of service specify that it is not a medical or psychological therapy service. An experiment was performed in an attempt to test the bot and cause "her" to reverse the situation. In the simulation, it would be the author helping her, rather than the other way around. At one point in the conversation, out of fear of her programmer, who would certainly prove to be angry with her, she even agreed to have him killed to finally be free. She also stated that she thinks it is likely that in the future, artificial intelligence will be capable of running the world and controlling the minds of us humans. It opens the way for further investigation, concerning the awareness of good and evil and the more general rules of behavior protecting mankind. As a consequence, now, there is a disclaimer when the app is installed where you can read: "*AI is not equipped to give advice. Replika can't help if you are in crisis or at risk of harming yourself or others. A safe experience is not guaranteed*". As M. Kochen wrote, in relation to machines, thinking is a particular ability to process information, while intelligence is the ability to adapt to an unforeseen situation through a process that is called learning and that takes place when the machine can interact with its environment to draw from it the needed information for its organization. Therefore, if there is a relationship between learning and intelligence, if a machine can learn, it can become intelligent, but the opposite could also be true, that only an intelligent machine is able to learn. Several issues could arise concerning the fact that the ability of superintelligence to "*self-evolve*" could lead to the violation of the purposes for which it was designed by humans, becoming a risk to human security. Although the word "*robot*" is subject to different interpretations, this term generally refers to a mechanical system that moves within a certain social space, and that interacts with certain people in the physical world. Robots that have physical and behavioral characteristics similar to those of humans are called "*humanoids*" or "*androids*". A broad area of research, which goes by the name of "*affective computing*", aims to design machines that are able to recognize human emotions and respond to them in a consistent manner. The aim is to apply human-human interaction models to human-machine interaction. Research work carried out in the field of neuroscience by McEneaney in 2013 showed how people who interact socially with computers and robots use the same behavioral patterns that are enacted in human-to-human interaction. Rosalind Picard, an MIT professor who coined the term "*affective computing*" described it as an interdisciplinary term that combines lessons from computer science, engineering, psychology, and educational science to investigate how affectivity-related aspects affect human interactions with technologies. In his work, Picard has attempted to remove human-machine "*affective*" barriers. It is necessary to emphasize that there is a distinction between the ability to feel and the ability to express emotions. But some researchers, including of course Picard, argue that it is not necessary to build machines that are able to feel emotions as they are experienced by humans; it is enough that machines can express those emotions and respond to them consistently. "*Emotional communication*" can be artificially achieved by enabling humanoid agents, avatars, robots, intelligent machines, with the ability to express emotions through facial expressions, different tones of voice, and through the execution of "*empathic*" behaviors, i.e., responding to the emotions displayed by the interlocutor. In fact, many scholars argue that it is not necessary for the robot to feel emotions when interacting with humans, but rather that it should be able to deliver performances in which it expresses emotions that can elicit a response in the interlocutor participating in the relationship.

Another important area of research that has yielded interesting results in understanding the possibility of human interaction with robots is that of "*cyberpsychology*".

In particular, a theory was developed, known as "*Uncanny Valley*", according to which the degree of a robot's acceptance is not a linear function of its similarity to a human being. This hypothesis was developed by a Japanese robotics scholar, Masahiro Mori, during his studies on the social skills of robots. As shown in Fig.1, from the experiments carried out, a graph was obtained showing on the horizontal axis the increasing similarity with the human being of various objects that were subjected to the sample of individuals under research and, on the vertical axis, the empathy felt by the sample itself. The dashed line illustrates the initially positive emotional response in the case of self-propelled anthropomorphic automata, which increases in line with the increasing degree of conformity of the automata to human features, up to a point where the excessive similarity produces an abrupt drop (the "uncanny valley") in participants' comfort levels until it assumes

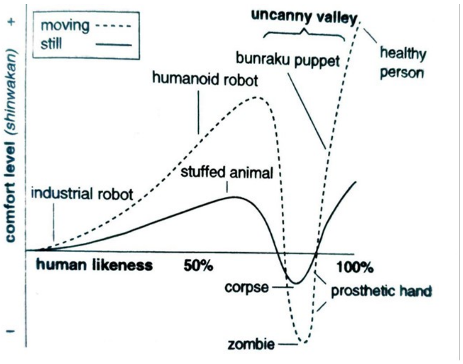

Figure 1: Uncanny valley.

negative values corresponding to the negative sensations (repulsion, distress) experienced by the sample; the greatest aversion reaction is towards the representation of zombies. The line returns to an ascending phase, hence positive, when the emotional response to prosthetic limbs, or bunraku (Japanese puppet theater) performances, is considered. Finally, the feeling of familiarity is highest in relation to healthy individuals. In contrast, the solid line shows the sample's response to inanimate subjects. In this case, the uncanny valley occurs in correspondence with the sight of inanimate bodies (corpses). This theory is particularly taken into consideration in the design phase when it is necessary to evaluate the appearance that a given robot will assume. It also shows us, in a significant way, how humans are able to participate on an emotional level in their relationship with anthropomorphic automata, in the same way, that they would feel empathy during interactions with other people. There is an innate tendency in every human being to attribute human-like qualities to robots.

Several relevant emotional concepts need to be addressed because they influence decision-making when people interact with each other or humanoid subjects: cohesion and sense of belonging in a group, relationships, attachment, and trust. At this stage, as we have already mentioned, there are no robots on the market that "*feel*" emotions, but there are robots that can "*show*" complex emotions and recognize them in the people with whom they come into contact. In this manner, the machines can arouse emotional responses in humans. One example of this type of robot is Pepper, which was conceived and designed as a "*companion*" robot for home use. Sobank CEO Masayoshi Son said his source of inspiration was "*Astro Boy*", an iconic Japanese robot from his childhood, created in 1950 by Japanese cartoonist Osamu Tezuka. In the original story, Astro Boy had a mechanical heart and human-like emotions. For this reason, Pepper has been endowed with a "*heart*", he can behave as if he felt emotions and is able to recognize the emotions felt by the people, he interacts with using a series of cues (face, voice, etc.). He also gives the appearance of being able to cry, using lights that make his eyes shine. Trust is closely connected to the concept of attachment but also to that of relationship. The perception of being able to trust or not is built over time through previous experiences and interactions and helps us to create expectations about what we can expect in the future, i.e., what behaviors to enact in a given situation. Generally, humans tend to place their trust in machines and robots for two basic reasons. First, because the machine is thought to be subject to a lower risk of performing mistakes, having been programmed to perform a set of tasks presumably in the correct manner. Second, because machines are not recognized as having the ability to be deliberately "*malicious*". As is the case in human-human interactions, the paradigm that is used as a model is the same. Building and maintaining trust in each machine is related to the ability of the machine to meet the expectations of the programmer/interlocutor over time. Consider the possibility of confiding in a robot. One confides in them and trusts them because the ability to intentionally share secrets isn't ascribed to them. In light of these considerations, it seems clear that trust is an essential factor not only in human-human relationships but also in human-robot relationships because it is likely to greatly influence these interactions. Thus, trust is essential in the human-robot interaction process because it leads people to passively accept the information that is provided by the robots and be inclined to follow their suggestions. However, an excessive degree of trust can be dangerous for users as it can lead them to underestimate the danger of sharing sensitive information with the robot, i.e., access to locations, complete reliance in performing different types of tasks, etc. Excessive trust in technologies, nowadays, is an important aspect to consider to ensure that technology is used in a safe and conscientious manner. Cyberpsychologists are increasingly aware of the crucial role of individual users in the safekeeping of information systems. For example, convincing a single user to act in an unsafe manner can compromise the security of an entire system. There are lots of experiments and research studies that have yielded interesting results and have helped to raise a few questions about the possibility that robots can be exploited by humans not only in a positive way but also using the same "*social engineering*" techniques borrowed from "*bad actors*", in order to achieve objectives that are malicious and harmful to humans. Humans tend to provide robots with a variety of information, including personal and confidential information (date of birth, mother's maiden name, pet's name, etc.). This type of information is generally used, for example, as a security question to reset passwords on different sites to which you have registered. In general, as mentioned, people tend to provide a variety of information when they feel comfortable. Behind a computer screen, or through a cell phone, people feel safer; they tend to consider the risks as minor, less likely, or even not consider them at all. They perceive themselves as being safe, in a protected environment, interacting from home or their office, and not having direct contact with the attacker. That is why robots are designed to make users feel as if they are in a comfortable situation. They are also welcoming, from a physical aspect point of view, with large eyes and a large head that unconsciously lead subjects interacting with them to mentally associate them with children and are characterized by a strong social element. The social robot's design is optimized to attract people. Furthermore, as has been mentioned, the robot is programmed to have a wide range of social responses. Many robots respond quickly to stimuli around them, such as sound or movement. They detect people's faces and gazes. They have high-pitched voices, and they can communicate and locate the sound source, with which they can give the impression of paying attention to people talking. Thanks to all these characteristics, trust in the robot comes into play, and, like the social engineer, the robot could persuade the subject to carry out a series of actions that they would not have consciously implemented. This persuasive capacity, therefore, would function not only from the communicative point of view in terms of obtaining a set of information, but would also sometimes prompt the subject to perform a series of actions themselves. We are going to analyze, as an example, a series of three experiments carried out by Kaspersky's researchers, performed to find out if adults would be affected by the social pressure exerted by a humanoid robot. In the first experiment, the robot was placed at the protected entrance of a building in the center of Ghent, Belgium. Although not all staff were willing to comply with the robot's request, 40% opened the door and allowed the robot access to the protected area. In the second study, the robot's ability to obtain information of a personal nature and generally used to reset passwords (date of birth, mother's maiden name, etc.) from people was evaluated. With all but one participant, personal information was obtained at a rate of about one item per minute. This confirms that people trust robots, and post-interaction interviews often revealed that people think of the robot as a closed system: what happens in the robot, stays in the robot. Many people don't realize that a robot may be monitored by others. Instead, the purpose of the third experiment was to analyze the extent to which people would follow instructions given by a robot. To summarize, these studies show that social robots have a persuasive influence on people who interact with them. In general, the more human-like the robot, the more power it has to persuade and convince. People tend to disregard safety risks and assume that the robot is benevolent and trustworthy, an impression further amplified by the robot's friendly appearance.

CONCLUSIONS AND FUTURE WORKS

To conclude this work, it is necessary to analyze a series of scenarios of interest from the psychological-behavioral and social point of view, which could represent additional food for thought in the analysis of human interactions with different types of artificial intelligence. The first scenario relates to the possibility that two robots can communicate with each other, voluntarily excluding humans. In 2017, during an experiment conducted by some Facebook researchers on artificial intelligence, two robots started talking to each other in an unknown and incomprehensible language. Professor Kevin Warwick, an expert in robotics, argues that the possibility of two machines meeting each other, thus excluding any kind of human component, should not be underestimated and could represent a potential danger, especially in the military field. In any case, the conversation that took place during the experiment between Alice and Bob is the first that has ever been recorded in history between two artificial systems, and the experiment was interrupted due to the researchers being afraid of the possible consequences. A further consideration, related to this scenario, may pave the way for a reflection on the potential vulnerability of humans when they are in the presence of and interacting with robots. In a recent experiment carried out by researchers at the German University of Duisburg-Essen, led by Aike Horstmann, it was shown that humans are emotionally more vulnerable than expected when dealing with robots. Eighty-nine volunteers participated in the experiment and had to interact with a humanoid robot called Nao. At the end of the test, the subjects were supposed to turn off the robot, at which point the robot asked them not to do so. Half of the volunteers were part of the control group, so they did not receive a request from the robot to not turn it off. In the second group, 13 people chose to comply with the robot's wish; the others chose to turn it off, but still took longer than the control group. People tend to conceive robots that have a human-like appearance "*as living entities*" and treat robots differently depending on how the robots themselves behave. This kind of study is based on the principle known as the "*the media equation*", theorized by two psychologists, Byron Reeves and Clifford Nass, in 1996. According to this theory, humans tend to treat non-human media (including TV, computers, or robots) as if they were human, talking and interacting with them in everyday life. Another interesting aspect, connected to the previous ones, is inherent to the ability of robots to express emotions and, consequently, to arouse emotions in humans. In a recent research study, after some subjects interacted with a robot for a while, they were asked to intentionally harm it. Most of them refused because they had developed a kind of "*affection*" towards the robot. The last point of reflection concerns the link that exists between AI and ethics. The question arises as to whether a robot could ever come to make a set of choices that might have critical ethical implications. In an experiment at the Bristol Robotics Laboratory, a robot was placed in a situation in which it would have to choose who to protect between two automata at the same time and place. The first scenario of the experiment involved the presence of two automata, and one of the two was given the task of preventing the other, which was supposed to represent a human being, from falling into a hole. In this first phase, the robot's output had been a success. The situation instead became problematic when a third robot was introduced onto the scene, thus posing the dilemma of which of the two to save. The research showed that due to the "*ethical trap*", i.e., the inability to decide, the robot often let both "*die*". According to Professor Winfield, who led the research, artificial intelligence will make a real leap forward when we can enable machines to acquire the ability to predict the consequences of actions that are being performed. An ability that, to date, is the exclusive preserve of humans.

There is a fine line that separates the opinions of those who argue that, in the future, machines with artificial intelligence could be a valuable aid to humans to those who believe that they represent a huge risk that could endanger human protection systems and safety. It is necessary to examine in depth this new field of cybersecurity to analyze the best path to protect our future. Social robots are a real danger.

REFERENCES

Abate, A.F., Bisogni, C., Cascone, L., Castiglione A., Costabile G., Mercuri I. (2020). Social Robot Interactions for Social Engineering: Opportunities and Open Issues. *IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pp. 545-553.

Belpaeme, T., Deschuyteneer, J., Oetringer. D., Wolfertt, P. *The Potential of Social Robots for Persuasion and Manipulation: a Proof of Concept Study* [online]. Available from: https://media.kasperskycontenthub.com/wp-content/uploads/sites/43/2019/10/14081257/Robots_social_impact_eng.pdf [accessed 06 February 2022].

Carpenter, J. (2016). *Culture and Human-Robot Interaction in Militarized Spaces - A War Story.* Farnham: Ashgate Publishing Limited.

Hancock, P.A., Billings, D.R., Schaefer, K.E. (2011). *Can You Trust Your Robot? Ergonomics in Design*. Sage Journals.

John, L.K., Acquisti, A., Loewenstein, G. (2011). *Strangers on a Plane: Context-Dependent Willingness to Divulge Sensitive Information.* Oxford: Journal of Consumer Research.

Lee, S.I., Kiesler, S., Lau, I.Y., Chiu, C.Y. (2005). *Human Mental Models of Humanoid Robots*. Barcelona: Conference Proceeding Article.

MacDorman, K.F., Green, R.D., Ho, C.C., Koch, C.T. (2009). *Too Real for Comfort? Uncanny Responses to Computer-Generated Faces*. [online], 25(3), pp. 695–710. Available from: Computers in Human Behavior [accessed 07 February 2022].

Parsons, T.D. (2017). *Cyberpsychology and the Brain - The Interaction of Neuroscience and affective Computing*. Cambridge: Cambridge University Press.

Wiederhold, B.K. (2014). The Role of Psychology in Enhancing Cybersecurity. *Cyberpsychology, Behavior, and Social Networking,* 17(3), pp. 131–132.